The Impact of Sections Headings on the Document Retrieval

Belkacem ABDELLI, Okba KAZAR[1], Jean-Marie PINON[2]
[1]University of Biskra
[2]LIRISLab, INSA de Lyon
France
{Bel_kac_em, kazarokba@yahoo.fr}, jean-marie.pinon@insa-lyon.fr

ABSTRACT: *With online publications, the current Web has become the largest source of digital documents, often stored in HTML, XML, PDF or DOC. Among the features of documents, note especially their logical structure, which represents their components such as chapters, sections, paragraphs, the document title, chapter titles, sections, etc...*

The section headings are meaningful; they are a good indicator of the content of paragraphs. For this reason we pay particular attention to these titles during the indexing process and research.

Our objective is to provide relevant access to digital documents, by the process of all sections titles to take advantage of their mining and importance in the research process. Experiments on a large corpus, INEX 2009 show effectiveness of our proposition an improvement in the precision of the results in IR.

Keywords: Information Retrieval, XML Logical Structure, Metadata, Mining

Received: 17 May 2014, Revised 20 June 2014, Accepted 26 June 2014

1.Introduction

Often long documents process many subjects, which are spread out in the logic sections of these documents, where each section has a title and paragraphs describing the content. The section headings have a major importance to indicate the content of paragraphs. They are very useful in the information retrieval

When the author, chose to put words in titles, and apply on these terms special police format, like a different size, readers understand that these terms are particularly important in the text. Here we can say that the author annotate terms that represent well the theme of the document.

When an author creates a hierarchy of title, where as a level n generalize the titles of a level $n + 1$, these amounts to achieve a segmentation of the document by the author. This segmentation can be very helpful in the process of analyzing and processing these documents and in the information retrieval domain.

Documentary searches systems added for each document data, such as title, author, date, abstract, keywords ... etc. they are called metadata. The use of metadata improves the result and minimizes irrelevant document and silence.

Systems that use metadata only have two major inconveniences; the first is, they do research on the metadata. The second is that they return the entire document, and not the parts, respond well to the request, which requires the user to read the entire document, (Abascal 2007).

For example, if a person is looking documents that deal with "*wordnet*", if that word does not appear explicitly in the metadata of a document, it will not be accepted as relevant documents.

Several studies in information retrieval domain, use the logical structure of documents to improve the precision and recall results, in particular XML RI, This work is limited to weight a few tags (title, abstract ...); or trying to calculate the weight of all the tags but they do not take into consideration the semantics hidden behind the "*title*" tag or the "*section title*" tag.

Our objective is to give special attention to the meaning of the "*title*" tag, and the "*section title*" tag, because these titles are a good indicator of the content of the document and paragraphs. In our work we use the titles of logical objects (chapters, sections, paragraphs) in the documents searching process.

For this we made a prototype that can extract and index the titles of logical objects of digital documents, and we tested it on a corpus INEX 2009. We build an inverted index, with several fields; one for main title of, the second forthe logical components titles, and one for text.

We aim to show that the use of a multi-field index improves the relevance of the results. Our prototype uses the Lucene[1], open source platform for indexing and retrieval of documents.

2. State of the Art

In what follows we will present works that has studied the importance of titles and the feasibility of their use

2.1 Works on the Importance of Titles

Among the studies that have investigated the importance of the section and sub-section headings in the documents, there are: [5] and [6]. The authors have shown that titles are important in two ways: first, as objects of logical organization of text used to segment, prioritize and structure the content of a document; secondly, they present the semantic content of documents, not explicitly, but as structured content, which allows players to build a "*mental model*" to understand the meaning of the text as and when it reads the document

This works shows the utility of titles in the classification and automatic extraction of relevant segments

2.2 Works Using Document Titles in IR

Little work has exploited the sections titles of documents, in information retrieval, most only operate the main title documents and ignore the other title that are in the content.

On the other side there are exist several studies in the web domain, that have used the titles and hyperlink (anchor text) that are in the HTML page, to improve information retrieval, Taking as an example the work of [8], who tried to extract title that found in the body of the web page and use theme in indexing the page, by proposing a new weighting method, inspired from Okapi BM25 method [14]. He showed that the use of these titles extracts can improve information retrieval.

Another work is the (Rapela 2001), which exploit some tags in HTML documents to improve the precision of the results of research; he affected the weights for the tag "*title*" [5], "*meta*" [4], "*link*" ... and the rest of the text the lowest weight.

[16] Proposes a method to retrieve scanned books, by exploiting their structure in indexing, where he creates an index, with multiple fields for each object of the book (table of contents, index ...). Using the research model BM25F [9], he showed that the results of recovery book improves with this method.

For focused or targeted search, [7] studied the impact of tags that represent the logical structure of documents represented

[1] http://lucene.apache.org/

in XML format, in focused search, to return the most relevant part (item) of XML document. It adds to each tag (title, abstract, section ...) weight to fully determine the relevant terms. [7] Also calculates the weight of all the tags and not just tag that contains the titles.

Work (Karen 2005) exploits the tree structure of XML documents to return the most relevant element of XML document, where the paths of nodes (tag) will be taken into account to calculate their weight

3. Proposed Approach

3.1 Context of the Work

Most XML documents are highly structured, they use the concept of tags to represent the content, and these tags are used not only to fragment the document elements, but to annotate the document in a way that we can understand the meaning of each tag: tag for logical structure, formatting tag, and tag for links…. The semantics of these tags can be very useful in information retrieval.

The following figure (Figure 1), we show an XML document, contains several tag which describes: the content, the format as (b, br ...), the structure (body: bdy, header: header; sec: section; st: section title ... ...), and tags for the links (link).

It is noted that each tag has a meaning, and describe very well the content of the document, some tag has a higher importance than other tag in the content description, as <*title*> tag that contains the main title of the document and <*st*> tag that indicates the content of a section title. These two tags summarize the contents of the document, and contain the most important terms in this document

```
< ? xml version = "1.0" encoding = "UTF- 8" ? >
……..
……
< title > 1956 in poetry  < / title >
< id > 5776006 < / id >
……..
……..
< bdy >
……..
< b > This is part of the  < linkxlink:type = "simple"
xlink:href = "../778 / 3327778.xml" >
List of years in poetry < / link > < / b >
……..
……..
< sec >
< st > Awards and honors < / st >
Consultant in Poetry to the Library of Congress……..

< / sec >
< / bdy >
…….
```

Figure 1. The tags in a wikipedia XML Document (1956 in poetry)

3.2 The Significance of the Structure (Tag) For RI

In this paper we propose an information retrieval system which takes into account the significance of the <*title*> tags, and <*section title*> to take their advantage in order to improve the precision of results research.

Most of the work that exploit the XML structure in the search for information, consider those two tags like other simple tags, and they will be indexed in the same way.

The idea is that an author when he wrote a paper designs a general view (the plan), and then begins to detail the various fragments (logical objects). The author proposes for each fragment (section, paragraph ...) document a title that reflects the content very well. Therefore documents have a logical structure based on hierarchy of titles.

We propose a method to extract all logical titles that are in the document to index them and use the semantics that is behind these titles.

Our approach is characterized by:

• The extraction of titles from the documents according to their logical hierarchy of objects

• The proposal of a new indexing method, where we have an inverted index with fields structure.

• In the research phase, the title fields will be used to return the most relevant documents general architecture of our approach

4. General Architecture of Our Approach

In this architecture, we have two main tasks, extracting titles from documents and index them, in title field, and index the rest of text content in another field. The result of the indexing task is an inverted index which contains two fields, one for words that are in section title and will be weighted by a heavy weight. The second field index words carried in the rest of the text.

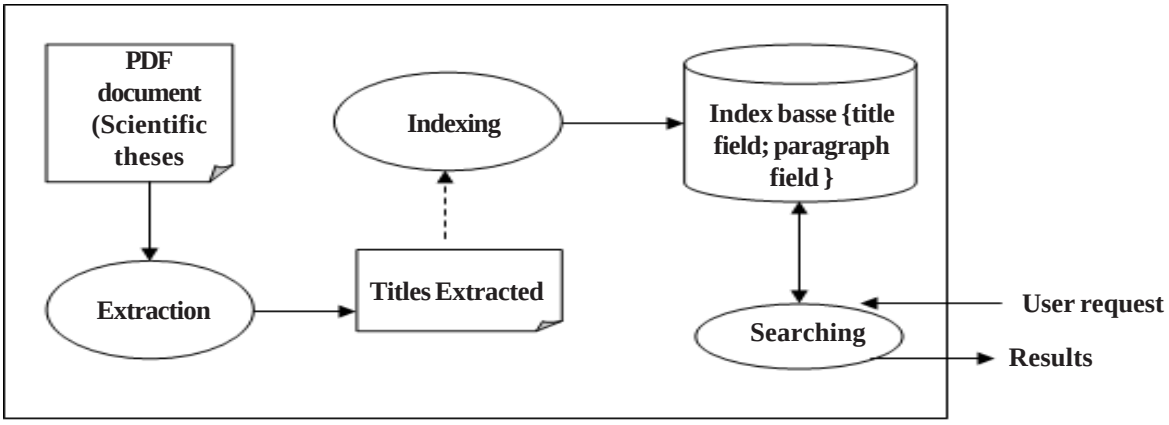

Figure 1. Exploitation of document titles in indexing

4.1 Extraction of Title

To extract all titles that are in the XML documents, we must browse all the nodes of this document. In the INEX corpus as the main document is still in the <*title*> tag in the header portion <*head*>, and <*st*> tag, if they exist, which contain the section titles are in the body of the document <*bdy*>. After extracting all titles, they will be sent to the indexing module to index them in the specified fields.

In the following table we show the number of terms in each field, and their size relative to the total size of the index, knowing that there are other fields represent the file name and path. Note that the corpus contains fewer terms of section titles compared with the main title and both are very small comparing to content field.

Field	Number of terms	Percentage
Content	3 889 525	37,6%
Principal title	790 045	7,64%
sections titles	297 340	2,87%

Table 1. The size of each field in the index

We note that the corpus contains fewer terms of section titles compared under the principal title. We can explain this by the fact that all documents in the corpus have a main title, but many of them do not have titles of sections

4.2 Indexing and Searching

Indexing is a set of processing operations on the documents to build an index to facilitate the search. The indexing steps are: extraction of words, then the weighting of these words, and finally the creation of the index.

In our work, we use the Lucene search engine to the Apache Foundation. [11]. Lucene can use the concept of fields in

indexing, where a field is a fragment of the document, as the document name, address of the document..... Lucene can also allow weight to a field

This structuring into Field, we appropriate, because in our work we will indexing titles of sections. Our inverted index will be composed of three fields:

- A «**document title»** field contains the general title of the document
- A «**section title»** field, which contains the result of the indexing terms of sections title,
- And the «**Body»** field that contains the result of indexing terms text of paragraph

Lucene represents the documents as the vector model (vector space model), introduced by [18]. In this model, documents are represented by a vector containing the indexing terms, the length of this vector is N (number of index terms in the collection)

• Stemming and Stop Words

Before indexing the words of the documents, we first begin by stemming phase, which eliminates the different variations of words and replace them with a single form (stem) to reduce the size of index (ex: research, searching becomes search). Another function is the removal of stop words like (the, it,).

To search word in a document, we use a simple model based on TFIDF relevance:

$$score(q,f) = coord(q,f) \cdot queryNorm(q) \cdot \sum_{t\ in\ q} (tf(t \text{ in } f)\ \ idf(t)^2 \cdot norm(t,f))$$

where: $tf(t \text{ in } f)$ is the frequency of the term t in the field *f* of the document;

$$tf(t\ in\ f) = \sqrt{freq}$$

$idf(t)$ represents the inverse of the term *t* frequency in the set of document;

$$idf(t) = log(numDocs / (docFreq + 1)) + 1$$

wherenumDocs represents the total number of documents in the corpus, and *docFreq* , the number of documents that contain the term *t*

coord (q, f) is a score factor based on the number of query terms contained in a specified field ; A field containing several query terms will have a higher score

$$coord(q, f) = tq / TQ$$

***tq* :** Numbers terms of the application that are in the field

***TQ* :** total number of query terms

queryNorm (q) is a normalizing factor used to make similar requests;

$$queryNorm(q) = 1 / \sqrt{(\sum idf\ (t)^2)}$$

norm (t, f) is used to standardize the size of fields (to make comparable fields) a shorter fields will have a higher score

$$norm(t, f) = 1 / \sqrt{\ (number\ of\ terms\ in\ the\ field)}$$

5. Validation

To validate our approach we realized prototype, which allows first step, extraction of titles from the documents and indexing and search these documents by exploiting the extracted titles.

Our prototype uses the open source Apache Lucene platform, to take advantage of its power in the indexing and searching text documents. We developed our prototype in the programming environment "*Eclipse*".

Our prototype consists of two algorithms: an algorithm without subtitles (content only) for research without the use of titles, and algorithm With Titles which use titles in information retrieval

5.1 Corpus

Our prototype is evaluated on the document collection of ad-hoc track of the Initiative for the Evaluation of XML retrieval (INEX) 2009 campaign; it is about 50.7 GB in size and with more than 2,600.000 documents. The documents contain a <*title*> tag to indicate the main title of the document and <*st*> tag to indicate the section title.

To evaluate our prototype we use queries (topics) (115 topics in INEX 2009) provided by the collection. The objective of using this collection is to enable the use of the assessment tool: *inex_eval*, which allows calculating the precision of defined recall points, *iP* [*x*] where *x* = {0.00, 0, 01; ..... 0.02; 100}

The evaluation in INEX privileges the precision than the recall, the ranking is based on the measurement *iP* [0.01]: precision at the point of recall 0.01 (Mathias 2012)

5.2 Evaluation

INEX companion provided a set of set of queries (topics) and for each query, judgments pertinences, which will used to evaluate our results

The metrics used in the INEX Ad hoc task (Kamps et al, 2008) are:

1. Precision interpolated four recall levels: [*r*] (*r* ∈ [0/00, 0.01, 0.05, 0.1]).

2.The mean of average interpolated precision *MAiP*: It is calculated as follows:

a. For a query r average interpolated precision *Aip* which measures the overall performance is calculated according to the 101 levels of recall ([0.00, 0.01, 0.02……., 1, 00]) :

$$Aip\,(r) = \frac{1}{101} \sum_{x = 0.00,\ 0.01,\ .....,\ 1.00} ip\,[x]$$

were *iP* [*x*] is the accuracy in the recall point *x*

b. MAiP is calculated as follows for *n* queries:

$$MAip = \frac{1}{n} \sum_{r = 1,\ 2,\ .....,\ n} Aip\,(r)$$

INEX uses the interpolated precision 1% recall (*iP* [0.01]) as the official measure. Our results are evaluated with the *inex_eval* assessment tool on the task Focused, The results show an improvement in *MAiP* and also in precision, for the first points of recall (*ip* [0.01]), when the title with the content is exploited in the search for information

5.3 Results

In our experiment, we try to show the effect of sections titles in the search for information, for this we created an index contains three fields (content, section titles and main title) and then launches research on these three fields.

We used the model TF-IDF Lucene. This model uses a normalization factor fields, a shorter field will have a higher score. And since deference size between fields (Table 1), we tried to decrease the gap between these fields by modifying this factor for each field:

After several trying, we found a value for the factor that can give good results, the value factor for the title field is: 5 (5 is the result of the division of the field size on content size field primarily: 37.6% / 7.64%), and 10 (37.6% / 2.87%) for the title field section

For The Content Field, No Change: norm (*t*, *f*) = 1 / √ (*nbr* term field)

For the main field Title: norm (*t*, *f*) = 1 / (5 * √ (*nbr* term field))

Titles for the field section: norm (*t*, *f*) = 1 / (10 * √ (*nbr* term field))

The following table shows the results obtained by running the search on a single field or multiple fields:

The table above shows the results, in one or more index fields. As already mentioned official INEX metrics are: (iP [0.01]), which represents precision, and (MAiP) which represents the recall.

Research in the Field	iP [0.01]	Improvement	MAiP	Improvement
Content	0.4494	-	0.13235	-
Principal title	0.5274	+17,35%	0.09925	-25 ,01%
Section titles	0.2979	-33,71	0.06264	-52,67%
Content and principal title	0.5713	**+27,12**	0.1338	+1,09%
Content and section titles	0.4966	+10,50	0.1440	**+8,80%**
Content and principal title and section titles	0.5850	**+30,70%**	0.14067	**+6,28%**

Table 2. Result of IR in the index fields

By comparing the different results with the results of research in the field content alone, we can see that:

• Research in the principal title only (principal Title field) showed a significant improvement, (17.35%) in precision (*iP* [0.01]), but a significant degradation in the recall (*MAiP*) of (-25 , 01 %)

• But research in the two fields together, principal Title with content, better improves precision (27.12) and also improves recall , but with a smaller value (1.09 %).

• Research in the section titles only (field section titles) degrades precision (-33.71) , and recall (-52.67 %).

• But research in the section titles and content together, shows improvement in precision (10.50), which remains as good as the precision of (principal title + content : 27.12) , and we also find some better improvement in recall (8.80 %)

• The research in three fields gives better precision (30.70 %) and a good reminder of (6.28 %) which is slightly less than (section titles + content)

The following table summarizes the comparison of the results with the results of research in the field (content only):

Research in the Field:	iP[0.00]	iP[0.01]	iP[0.05]	iP[0.10]	MAiP
Content	0.4711	0.4494	0.3940	0.3417	0.13235
Principal title	0.5642	0.5274	0.3842	0.2949	0.09925
Section titles	0.3423	0.2979	0.2198	0.1787	0.06264
Content and principal title	0.5970	**0.5713**	0.4700	0.3853	0.1338
Content and section titles	0.5087	0.4966	0.4400	0.3806	**0.1440**
Content and principal title and section titles	**0.6005**	**0.5850**	**0.4911**	**0.4001**	**0.14067**

Table 3. Comparison of the results of the field with the content field

Appraisal

• The search operator, the main title alone, or section titles does not improve the search, we must exploit upstream with the content of documents

• It can be concluded that the main title gives better precision in the top results

• And section titles improve the recall.

To understand our method, we present the curves of precision at different points of recall, we obtained as results of our experiments.

Figure 2; shows that the document search by exploiting their main title gives better accuracy in results returned first, but it

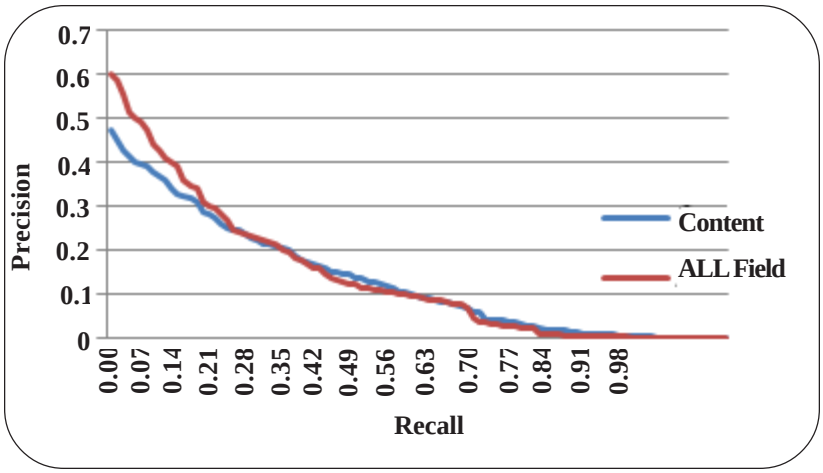

Figure 2. Precision at different recall points for research in the content alone, and content + principal

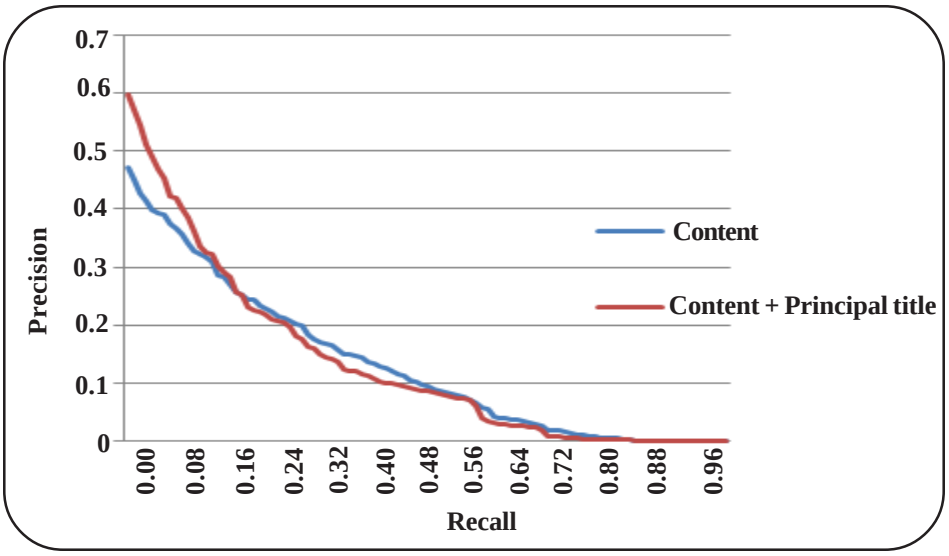

Figure 3. Precision at different recall points for research in the content alone, content + section titles

becomes slightly worse than looking in the content only from the point of recall(0.20).

5.4 Discussion

The number of section title in the corpus is too low (2.87% of the index), which is smaller than the number of primary shares (7.64%). We can explain this by the fact that all documents in the corpus have a main title, but many of them do not have titles of sections.

Despite this low rate, but the operation of section titles in information retrieval, shows an improvement in the precision and recall results.

6.conclusion

We propose in our work taking into account the shares of logical objects that are in the document, in the search for information. We made a prototype that can extract titles in documents and then use these titles in the indexing phase and the phase of the research information.

By comparing the results obtained when the titles and those obtained by not exploiting the titles we operate, we see that the results we obtained with the use of titles are better.

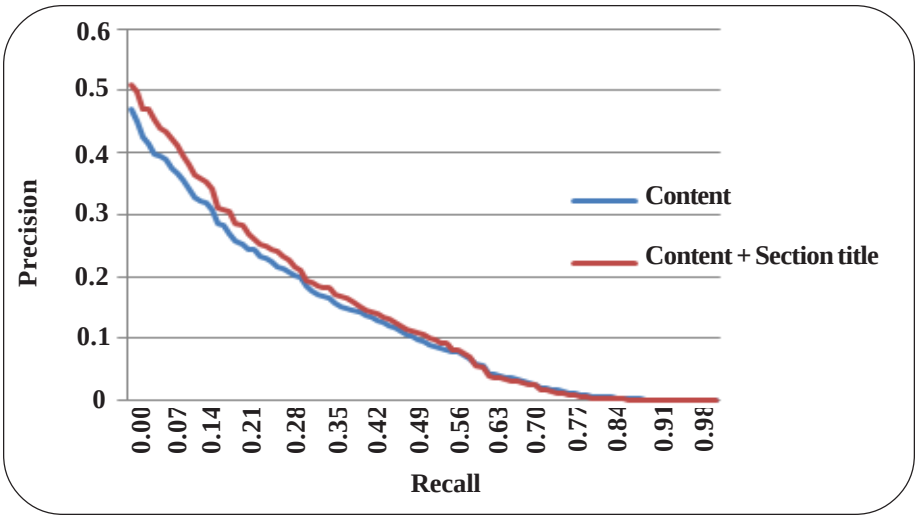

Figure 4. Precision at different points in return for research in the content alone, and content + title + primary section titles

One perspective to our work is to exploit the hierarchy of Section headings for return the most relevant element of the XML document, and not the entire document

A second perspective is to improve the quality of indexing using domain ontology in order to take into account the semantics. This will establish the relationship between the terms, eg (" *semi- structured document* " and "XML «) or (" *semantic resource* " and " *thesaurus* ").

References

[1] Abascal, R., Rumpler, B. (2007). Accès au contenu des thèses numériques par leur structure sémantique, Lavoisier Document numérique. 10, p 9-35.

[2] Salton, G., Wong, A., Yang, C. S. (1975). A vector space model for automatic indexing, *Communications of the ACM*, November. 18 (11) 613-620.

[3] Hervé, D., (2009). Jean-Luc M On tables of contents and how to recognize them *International Journal of Document Analysis and Recognition* (IJDAR), May 12, (1), p 1-20.

[4] Ho-Dac, L. -M., Jacques, M. -P., Rebeyrolle. (2004). Sur la fonction discursive des titres In L'unité texte S. Porhiel, Klingler, D., (Eds), Pleyben, Perspectives, p. 125-152.

[5] Jacques, M. -P., Rebeyrolle, J. (2006). Titres et structuration des documents », International Symposium: Discourse and Document, Caen (France), 15-16 juin, p1-12.

[6] Jean-Pierre, C., Boris, C., Hervé, D., et al. (2005). From Legacy Documents to XML: A Conversion Framework, Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, 3652, p 92-103.

[7] Karen, S., Mohand, B., Claude, C. (2006). Answering content and structure-based queries on XML documents using relevance propagation», *Information Systems*, 31, p 621–635.

[8] Lalmas, M.(2009). XML Information Retrieval , Encyclopedia of Library and Information Sciences, Bates, J., Maack, M. N. (Eds).

[9] Mathias, G., Largeron, C. (2012). BM25t: a BM25 extension for focused information retrieval, Knowledge and information systems, Springer, 32, p 217-241

[10] Stephen, R., Hugo, Z. (2009). The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval, April . 3 (4) 333-389.

[11] Xue, Y., Hu, Y., Xin, G. et al .(2007).Web page title extraction and its application, *Information Processing & Management*, September, 43 (5) 1332–1347.